# Minimum-sized Influential Node Set Selection for Social Networks under the Independent Cascade Model

Jing (Selena) He
Department of Computer Science
Kennesaw State University
Kennesaw, GA, 30144
jhe4@kennesaw.edu

Shouling Ji
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30308
sji@gatech.edu

Raheem Beyah
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30308
rbeyah@ece.gatech.edu

Zhipeng Cai
Department of Computer Science
Georgia State University
Atlanta, GA, 30303
zcai@cs.gsu.edu

## ABSTRACT

Social networks are important mediums for communication, information dissemination, and influence spreading. Most of existing works focus on understanding the characteristics of social networks or spreading information through the "word of mouth" effect of social networks. However, motivated by applications of alleviating social problems, such as drinking, smoking, addicting to gaming, and influence spreading problems, such as promoting new products, we propose a new optimization problem named the *Minimum-sized Influential Node Set* (MINS) selection problem, which is to identify the minimum-sized set of influential nodes, such that every node in the network could be influenced by these selected nodes no less than a threshold $\tau$. Our contributions are threefold. First, we prove that, under the independent cascade model, MINS is NP-hard. Subsequently, we present a greedy approximation algorithm to address the MINS selection problem. Moreover, the performance ratio of the greedy algorithm is analyzed. Finally, to validate the proposed greedy algorithm, extensive experiments and simulations are conducted both on real world coauthor data sets and random graphs.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Theory


*MobiHoc'14,* August 11–14, 2014, Philadelphia, PA, USA.

http://dx.doi.org/10.1145/2632951.2632975.

## Keywords

Viral Marketing; Influence Spreading; Social Networks; Minimum-sized Influential Node Set; Greedy Algorithm; NP-hard Problem; Performance Ratio

## 1. INTRODUCTION

A social network (*e.g.*, Facebook and MySpace) is a network made up of a set of nodes (such as individuals or organizations) and the social ties (such as relations and interactions) among these nodes. Ever since social networks appear, they play a fundamental role as a medium for spreading information, ideas, and influences among individuals. With the emergence of social applications (such as MSN, Wikis, Netflix, and Twitter), there have been tremendous interests in exploring social influences from individual-to-individual and individual-to-group interactions [1, 2, 3, 4]. This is because that the social influence is becoming a critical, complex, and subtle force to dominate the dynamics of a social network. As a result, extensive researches have been dedicated to select a set of influential users to spread ideas and information within a group recently. In this paper, motivated by influence spreading applications and alleviating social problems, such as drinking, smoking, and addicting to gaming [5], we aim to find a Minimum-sized Influential Node Set (MINS), which influences every individual in a social network no less than a pre-defined threshold $\tau$.

Constructing an MINS is helpful to alleviate the aforementioned social problems, and it is also helpful for promoting new products in a social network. Consider the following scenario as a motivation example. A small company wants to market a new product in a community. For saving the budget, but to get the maximum profit, the company would like to distribute the product samples to a small number of initial influential users in the community. The company wishes that these initial users would love the product and start to influence their friends in the community. The goal is that every user in the community is influenced by the initially selected users no less than $\tau$ eventually. To sum up, the specific problem we investigate in this paper is the following: given a social network, and a threshold $\tau$, how to identify a minimum-sized subset of the individuals in the network such

that the subset can result in a influence on every individual in the social network no less than $\tau$.

The proposed MINS selection problem is different from the *influence maximization* problem [6], which tries to find a subset of individuals $M$, such that, for a preset threshold $k$, $|M| = k$ and $M$ has the maximum expected number of influenced individuals over other subsets of size $k$. The problem investigated in this paper is to find a minimum-sized set of influential nodes, so that they have the influence on every node in the network no less than $\tau$. A related work to our research is [7], which is to find a minimum-sized *Positive Influence Dominating Set* (PIDS) $D$, so that every node has at least half of its neighbors in $D$. Actually, the authors in [7] studied the MINS selection problem under the *deterministic linear threshold model*, in which the influence from a pair of nodes is represented by a weight and an individual can be positively influenced when the sum of the weights exceeds a pre-determined threshold. To be specific, the authors in [7] assumed that the influence of each social tie is always 1, and an individual can be positively influenced when at least half of its neighbor nodes are in $D$. Nevertheless, the deterministic linear threshold model cannot fully characterize the social influence between each pair of nodes in a real social network. This is because that, in the physical world, the strength of the social influence between different pairs of nodes may be different and is actually a probabilistic value [1, 2, 3, 4]. Hence, we explore the MINS selection problem under the *independent cascade model* [6], where individuals can influence their neighbors with certain independent probabilities. In this paper, we study MINS under the independent cascade model. We first analyze the NP-hardness of MINS, and then propose a greedy approximation algorithm to solve the problem with performance analysis. Particularly, the main contributions of this paper are summarized as follows.

1) We introduce a new optimization problem, named the Minimum-sized Influential Node Set (MINS) selection problem, which is to identify the minimum-sized set of influential nodes, that could influence every node in the network no less than a pre-defined threshold $\tau$. We further prove that it is a NP-hard problem under the independent cascade model.

2) We define a polymatroid contribution function, which suggests us a greedy approximation algorithm called MINS-GREEDY to address the MINS selection problem. Comprehensive theoretical analysis about its performance ratio is also given in the paper.

3) We also conduct extensive experiments and simulations to validate our proposed algorithm both on real world coauthor data sets and random graphs. The experiment and simulation results show that the proposed greedy algorithm works well to solve the MINS selection problem. More importantly, the solutions obtained by the greedy algorithm is very close to the optimal solutions of MINS in small scale networks.

The rest of this paper is organized as follows: in Section 2, we review some related literatures. In Section 3, we first introduce the network model and then we formally define the MINS selection problem and prove its NP-hardness. The greedy algorithm and the theoretical analysis of the algorithm are presented in Section 4. The experimental results are presented in Section 5 to validate our proposed algorithm. Finally, the paper is concluded in Section 6.

# 2. RELATED WORK

In this section, we first briefly review the related works of the influence maximization problem, and the PIDS problem. Subsequently, we summarize some related literatures of social influence analysis, followed by some remarks.

## 2.1 Influence Maximization Problem

Domingos *et al.* [8, 9] were the first to emphasize the node selection problem when propagating information through social networks. They considered the social relations of individuals and proposed a probabilistic information propagation model for the problem, as well as several heuristic solutions. Subsequently, Kempe *et al.* formulated the influence maximization problem and studied the problem under two different models *i.e.*, the linear threshold model and the independent cascade model in [6, 10]. They proposed greedy algorithms and analyzed their performance ratios, which are $1-\frac{1}{e}$ under both models. To address the scalability problem of the algorithms in [6, 10], Leskovec *et al.* [11] presented a "lazy-forward" optimization scheme in selecting initial nodes, which greatly reduces the number of influence spread evaluations. Laterly, Chen *et al.* [12, 13, 14] showed that the problem of computing exact influence in social networks under both the linear threshold model and the independent cascade model is #P-Hard. They also proposed scalable algorithms under both models, which are faster than the greedy algorithms proposed in [6, 10].

On the other hand, Goyal *et al.* [15] studied the influence maximization problem from the data-based perspective. They introduced a new model called *credit distribution*, which directly leverages available propagation traces to learn how influence flows in the network and adopt it to estimate the expected influence spread. They also showed that the influence maximization problem under the credit distribution model is NP-hard, and an approximation algorithm is designed. Dinh *et al.* [16] investigated the cost-effective massive viral marketing problem, taking into the consideration the limited influence propagation. They proposed an efficient algorithm called VirAds to tackle the viral marketing problem on large-scale networks. VirAds guarantees a relative error bound of $O(1)$ from the optimal solutions in power-law networks. Zou *et al.* were the first to add the latency constraint to the influence maximization problem under the linear threshold model, called the *fast information propagation* problem in [17]. They further proved that the fast information propagation problem is NP-hard in [18]. Moreover, two heuristic algorithms are given and their performance ratios are also analyzed. He *et al.* explored the influence maximization problem considering both positive influence and negative influence in [19].

## 2.2 Positive Influence Dominating Set Problem

Wang *et al.* first proposed the Positive Influence Dominating Set (PIDS) problem under the deterministic linear threshold model in [5], which is to find a set of nodes $D$ such that every node in the network has at least half of its neighbor nodes in $D$. They proposed a selection algorithm and analyzed its performance on a real online social network data set. Subsequently, Zhu *et al.* proved that PIDS is APX-hard and proposed two greedy algorithms with approximation ratio analysis in [7] and [20], respectively.

## 2.3 Social Influence Analysis

Saito *et al.* predicted the information diffusion probabilities under the independent cascade model in [1]. They formally defined the likelihood maximization problem and then applied an Expectation-Maximization (EM) algorithm to solve it. Subsequently, Tang *et al.* argued that the effect of the social influence from different angles (topics) may be different. Hence, they introduced Topical Affinity Propagation (TAP) to model topic-related social influence in large social networks in [2]. Later, Tang *et al.* [4] proposed a Dynamic Factor Graph (DFG) model to incorporate the time information to analyze dynamic social influences. Similarly, Goyal *et al.* [3] studied the problem of learning the influence probabilities from historical node actions.

## 2.4 Remarks

All the above mentioned existing works fall into three categories: one is to understand the properties and characteristics of social networks. Another is to study the influence maximization problem with or without the time constraint. The last is the PIDS problem. Our work different from the influence maximization problem is that we find a minimum-sized set of individuals that guarantees the influence on every node in the network no less than a threshold $\tau$, while the influence maximization problem focuses on choosing a subset of a pre-established size $k$ that maximizes the spread of information. Moreover, our work is also different from the PIDS problem. We study the MINS selection problem under the independent cascade model which is more practical, while PIDS is investigated under the deterministic linear threshold model.

# 3. PROBLEM DEFINITION AND HARDNESS ANALYSIS

In this section, we first introduce the network model. Subsequently, we formally define the MINS selection problem and make some remarks for the proposed problem. Finally, we analyze the hardness of the MINS selection problem.

## 3.1 Network Model

We model a social network by an undirected graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, where $\mathbb{V}$ is the set of $n$ nodes, denoted by $u_i$, and $0 \leq i < n$. $i$ is called the node ID of $u_i$. An undirected edge $(u_i, u_j) \in \mathbb{E}$ represents a social tie between the pair of nodes. $\Lambda(\mathbb{E}) = \{p_{ij} \mid \text{if } (u_i, u_j) \in \mathbb{E}, 0 < p_{ij} \leq 1, \text{ else } p_{ij} = 0\}$, where $p_{ij}$ indicates the social influence between nodes $u_i$ and $u_j$ [1]. For simplicity, we assume the links are undirected (bidirectional), which means two linked nodes have the same social influence (*i.e.*, $p_{ij}$ value) on each other. Additionally, we assume every node has an edge to itself. The corresponding $p_{ii}$ value ($0 \leq p_{ii} \leq 1$) represents the social influence to itself.

## 3.2 Problem Definition

The objective of the MINS selection problem is to identify a subset of influential nodes as the initialized nodes. Such that, all the other nodes in a social network can be influenced by these nodes no less than a threshold $\tau$. For convenient, we call the initial nodes been selected as *active nodes*, otherwise, *inactive nodes*. Therefore, how to define *influence* is critical to solve the MINS selection problem. In the following, we first formally define some terminologies, and then give the definition of the MINS selection problem.

[1]This model is reasonable since many empirical studies have analyzed the social influence probabilities between nodes [1, 2, 3, 4].

DEFINITION 3.1. *Influential Node Set ($\mathbb{I}$):* For social network $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, the influential node set is a subset $\mathbb{I} \subseteq \mathbb{V}$, such that all the nodes in $\mathbb{I}$ are initially selected to be the active nodes.

DEFINITION 3.2. *Active Neighbor Set ($\mathbb{A}^{\mathbb{I}}(u_i)$):* For social network $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, $\forall u_i \in \mathbb{V}$, the active neighbor set of $u_i$ is defined as:

$$\mathbb{A}^{\mathbb{I}}(u_i) = \{u_i\} \cup \{u_j \mid (u_i, u_j) \in \mathbb{E}, u_j \in \mathbb{I}\}.$$

Followed by Definition 3.2, we know that the set $\mathbb{A}^{\mathbb{I}}(u_i)$ includes all the active neighbor nodes of $u_i$ and $u_i$ itself. Since every node has a self-circled edge to itself as defined in Section 3.1, we define the *self influence* as follows:

DEFINITION 3.3. *Self Influence ($p_{ii}$):* For social network $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, and $\forall u_i \in \mathbb{V}$, the self influence of $u_i$ is defined as: $p_{ii} = 1$, if $u_i \in \mathbb{I}$; otherwise, $p_{ii} = 0$.

DEFINITION 3.4. *Influence:* For social network $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, a node $u_i \in \mathbb{V}$, and an influential node set $\mathbb{I}$, we define a joint influence probability of $\mathbb{A}^{\mathbb{I}}(u_i)$ on $u_i$, denoted by $p_{u_i}(\mathbb{A}^{\mathbb{I}}(u_i))$ as

$$p_{u_i}(\mathbb{A}^{\mathbb{I}}(u_i)) = 1 - \prod_{u_j \in \mathbb{A}^{\mathbb{I}}(u_i)} (1 - p_{ij}).$$

If $p_{u_i}(\mathbb{A}^{\mathbb{I}}(u_i)) \geq \tau$, where $0 < \tau < 1$ is a pre-defined threshold, then $u_i$ is said been influenced. Otherwise, $u_i$ is not been influenced.

DEFINITION 3.5. *Minimum-sized Influential Node Set (MINS).* For social network $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, the MINS selection problem is to find a minimum-sized influential node set $\mathbb{I} \subseteq \mathbb{V}$, such that $\forall u_i \in \mathbb{V}$, $u_i$ is influenced, *i.e.*, $p_{u_i}(\mathbb{A}^{\mathbb{I}}(u_i)) = 1 - \prod_{u_j \in \mathbb{A}^{\mathbb{I}}(u_i)} (1 - p_{ij}) \geq \tau$.

In this paper, we study the MINS selection problem. First, we analyze the complexity of the problem, which is NP-hard. Subsequently, we propose a greedy algorithm called MINS-GREEDY to solve the problem with performance analysis.

## 3.3 Problem Hardness Analysis

In general, given an arbitrary threshold $\tau$, the MINS selection problem is NP-hard. We prove the complexity of the MINS selection problem in a general graph by constructing a polynomial reduction from the Vertex Cover (VC) problem to MINS as shown in the following theorem. We only provide proof sketch of Theorem 1 due to space limitation.

THEOREM 1. The MINS selection problem is NP-hard.

*Proof Sketch:* We prove the NP-hardness of MINS by constructing a polynomial-time many-one reduction which converts instances of one decision problem, *i.e.*, the decision version of the well known Vertex Cover (VC) problem into instances of a second decision problem, *i.e.*, the decision version of a specific case of the MINS selection problem.

The decision version of the VC problem is defined as follows: given a graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, where $\Lambda(\mathbb{E}) = \{1 \mid (u_i, u_j) \in \mathbb{E}; u_i, u_j \in \mathbb{V}\}$, and a positive integer $d$, determine whether $\mathbb{G}$ has a vertex cover[2] of size at most $d$.

The decision version of a specific case of the MINS selection problem is: given a graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, where $\Lambda(\mathbb{E}) = \{p \mid (u_i, u_j) \in \mathbb{E}, 0 < p < \tau, u_i, u_j \in \mathbb{V}\}$[3], a pre-defined threshold $0 < \tau < 1$, and a positive integer $k$, determine whether there exists an influential node set $\mathbb{I} \subseteq \mathbb{V}$, such that the size of $\mathbb{I}$ is $|\mathbb{I}| \leq k$, and every node in $\mathbb{G}$ is influenced, *i.e.*,

$$\forall u_i \in \mathbb{V}, p_{u_i}(\mathbb{A}^{\mathbb{I}}(u_i)) = 1 - \prod_{u_j \in \mathbb{A}^{\mathbb{I}}(u_i)} (1 - p_{ij}) \geq \tau.$$

First, for an instance of the VC problem, denoted by graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, where $\Lambda(\mathbb{E}) = \{1 \mid (u_i, u_j) \in \mathbb{E}; u_i, u_j \in \mathbb{V}\}$, we construct a new graph $\widehat{\mathbb{G}}$ as follows:

1) We create $|\mathbb{V}| + |\mathbb{E}|$ nodes with $|\mathbb{V}|$ nodes $\{v_{u_1}, v_{u_2}, \cdots, v_{u_{|\mathbb{V}|}}\}$ representing the nodes in $\mathbb{G}$ and $|\mathbb{E}|$ nodes $\{v_{e_1}, v_{e_2}, \cdots, v_{e_{|\mathbb{E}|}}\}$ representing the edges in $\mathbb{G}$.

2) We add an edge with influence weight $p$ between nodes $v_{u_i}$ and $v_{e_j}$ if and only if node $u_i$ is an endpoint of edge $e_j$.

3) We attach additional $\lceil \log_{1-p}(1-\tau) \rceil$ nodes to each node $v_{u_i}$, denoted by set $\mathbf{v}^a_{u_i} = \{v^j_{u_i} \mid 1 \leq j \leq \lceil \log_{1-p}(1-\tau) \rceil\}$.

4) We attach additional $\lceil \log_{1-p}(1-\tau) \rceil - 1$ nodes to each node $v_{e_j}$, denoted by set $\mathbf{v}^a_{e_j} = \{v^j_{e_j} \mid 1 \leq j \leq \lceil \log_{1-p}(1-\tau) \rceil - 1\}$.

5) Then, we have $\widehat{\mathbb{G}} = \{\widehat{\mathbb{V}}, \widehat{\mathbb{E}}\}$, where $\widehat{\mathbb{V}} = \{v_{u_1}, \cdots, v_{u_{|\mathbb{V}|}}\} \cup \{v_{e_1}, \cdots, v_{e_{|\mathbb{E}|}}\} \cup \bigcup_{i=1}^{|\mathbb{V}|} \mathbf{v}^a_{u_i} \cup \bigcup_{i=1}^{|\mathbb{E}|} \mathbf{v}^a_{e_i}$, $\widehat{\mathbb{E}}$ is the set of all the edges associated with the nodes in $\widehat{\mathbb{V}}$, and $\Lambda(\widehat{\mathbb{E}}) = \{p \mid \text{for every edge in } \widehat{\mathbb{E}}\}$.

Taking the network shown in Fig. 1(a) as an example to illustrate the construction procedure from $\mathbb{G}$ to $\widehat{\mathbb{G}}$. There are 4 nodes and 6 edges in $\mathbb{G}$. Therefore, we first create $\{v_{u_i}\}_{i=1}^4$ and $\{v_{e_j}\}_{j=1}^6$ nodes in $\widehat{\mathbb{G}}$. And then we add edges with influence weight $p$ between nodes $v_{u_i}$ and $v_{e_j}$ based on the topology shown in $\mathbb{G}$. Subsequently, we add additional nodes $\mathbf{v}^a_{u_i} = \{v^j_{u_i} \mid 1 \leq j \leq \lceil \log_{1-p}(1-\tau) \rceil\}$ to each node $v_{u_i}$. Similarly, we add additional nodes $\mathbf{v}^a_{e_j} = \{v^j_{e_j} \mid 1 \leq j \leq \lceil \log_{1-p}(1-\tau) \rceil - 1\}$ to each node $v_{e_j}$. The influence weights on all the additional edges are $p$. Finally, the new graph $\widehat{\mathbb{G}}$ is constructed as shown in Fig. 1(b).

We will prove that $\mathbb{G}$ has a VC $\mathbb{D}$ of size at most $d$ if and only if $\widehat{\mathbb{G}}$ has an influential node set $\mathbb{I}$ of size at most $k$ by setting $k = |\mathbb{V}|\lceil \log_{1-p}(1-\tau) \rceil + |\mathbb{E}|(\lceil \log_{1-p}(1-\tau) \rceil - 1) + d$.

In conclusion, we proved that a specific case of the MINS selection problem is NP-hard, since the VC problem is NP-hard. Consequently, the general MINS selection problem is also at least NP-hard. □

[2]A vertex cover is defined as a subset of nodes in a graph $\mathbb{G}$ such that each edge of the graph is incident to at least one vertex of the set.

[3]The graph model is a special case of the network model defined in Section 3.1.

Based on Theorem 1, we conclude that MINS cannot be solved in polynomial time unless P = NP. Therefore, we proposed a greedy algorithm to solve the problem in the next section.

# 4. GREEDY ALGORITHM

Since MINS is NP-hard, we propose a greedy algorithm to solve it. The greedy criterion is that the node influencing the most other nodes will be added into MINS first, which is defined by the following contribution function:

DEFINITION 4.1. *Contribution function ($f(\mathbb{I})$).* For a social network represented by graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$, and an influential node set $\mathbb{I}$, the contribution function of $\mathbb{I}$ to $\mathbb{G}$ is defined as:

$$f(\mathbb{I}) = -\theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho^{\mathbb{I}}_{u_i}, 1-\tau)],$$

where $\varrho^{\mathbb{I}}_{u_i} = \prod_{u_j \in \mathbb{A}^{\mathbb{I}}(u_i)} (1 - p_{ij})$, $\theta = \max(1/c_1, 1/c_2, 1/c_3)$, $c_1 = -\log(1-\tau)$, $c_2 = \log(\frac{1}{\max_{p_{ij}<1}(1-p_{ij})})$, and $c_3 = \log(\frac{\min\{\prod_{u_j \in \mathbb{S}}(1-p_{ij}) \mid u_j \in \mathbb{V}, \mathbb{S} \subseteq \mathbb{V}, \prod_{u_j \in \mathbb{S}}(1-p_{ij}) > 1-\tau\}}{1-\tau})$.

For the defined contribution function, it has some important properties as shown in the following lemma. The proof of Lemma 1 is omitted due to space limitation.

LEMMA 1. *1)* $f(\emptyset) = 0$. *2)* $f(\mathbb{I})$ is an increasing function.

Based on the defined contribution function, we propose a greedy algorithm called MINS-GREEDY as shown in Algorithm 1. MINS-GREEDY starts from an empty influential node set $\mathbb{I}$. Each time, it adds the node having the maximum $f(\cdot)$ value into $\mathbb{I}$. The algorithm terminates when $f(\mathbb{I}) = -\theta * \log[(1-\tau)^{|\mathbb{V}|}]$.

**Algorithm 1: MINS-GREEDY Algorithm**

**Input**: A social network represented by graph $\mathbb{G}(\mathbb{V}, \mathbb{E}, \Lambda(\mathbb{E}))$; a pre-defined threshold $\tau$

```
I = ∅;
while f(I) < −θ * log[(1 − τ)^|V|] do
    choose u ∈ V \ I to maximize f(I ⋃ {u});
    I = I ⋃ {u};
return I;
```

To better understand Algorithm 1, we use the social network represented by the graph shown in Fig. 2(a) to illustrate the selection procedure as follows. In the example, $\theta = 4.5$.

1) First round: $\mathbb{I} = \emptyset$.

2) Second round: we first compute $f(\mathbb{I} = \{u_1\}) = 8.64$, $f(\mathbb{I} = \{u_2\}) = 8.64$, $f(\mathbb{I} = \{u_3\}) = 9.63$, $f(\mathbb{I} = \{u_4\}) = 11.21$. Therefore, we have $\mathbb{I} = \{u_4\}$, which has the maximum $f(\mathbb{I})$ value. However, $f(\mathbb{I} = \{u_4\}) = 11.21 < -4.5 * \log(0.2^4) = 12.60$. Consequently, the selection procedure continues.

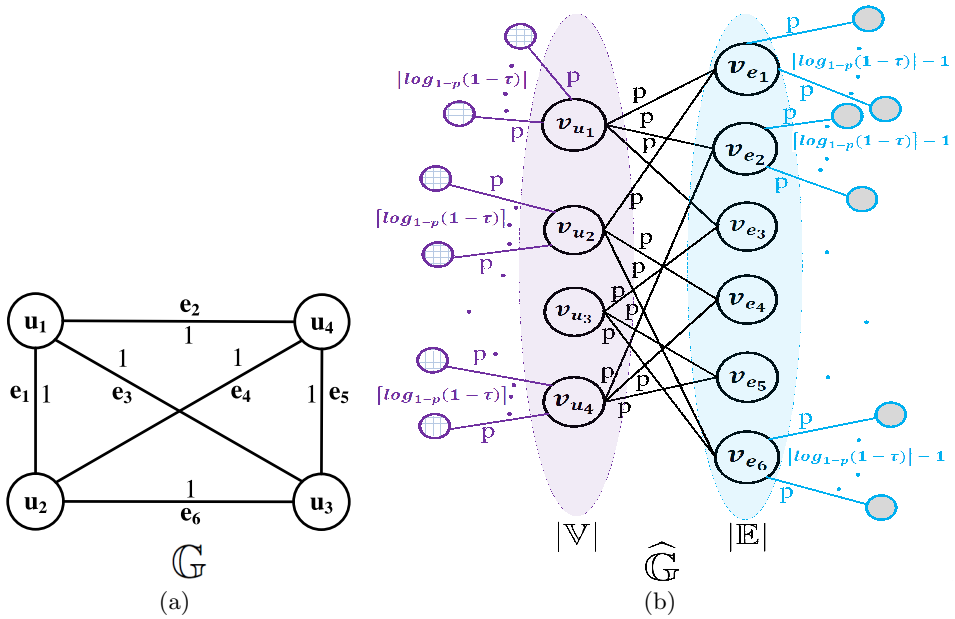



**Figure 1: Illustration of the construction from $\mathbb{G}$ to $\widehat{\mathbb{G}}$.**

3) Third round: we first compute $f(\mathbb{I} = \{u_1, u_4\}) = 12.60$, $f(\mathbb{I} = \{u_2, u_4\}) = 12.60$, $f(\mathbb{I} = \{u_3, u_4\}) = 12.60$. Therefore, we have $\mathbb{I} = \{u_1, u_4\}$[4]. Since $f(\mathbb{I} = \{u_1, u_4\}) = -4.5 * \log(0.2^4) = 12.60$, algorithm terminates and outputs set $\mathbb{I} = \{u_1, u_4\}$ as shown in Fig. 2(b), where black nodes represent the selected influential nodes.

It is easy to check that $u_2$ and $u_3$ are both influenced. Hence, the constructed $\mathbb{I}$ by running Algorithm 1 is a feasible solution for the MINS selection problem. From this example, we know that the time complexity of Algorithm 1 is $O(n^2)$ in worst case.

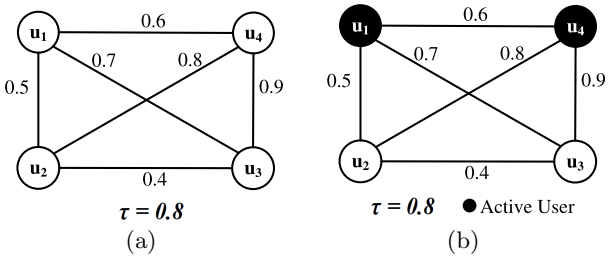



**Figure 2: Illustration of MINS-Greedy algorithm.**

Now, we theoretically show the correctness of Algorithm 1 in the following theorem. The proof of Theorem 2 is omitted due to space limitation.

THEOREM 2. Algorithm 1 produces a feasible solution of the MINS selection problem. To be specific, *1)* Algorithm 1 terminates for sure. *2)* $f(\mathbb{I}) = -\lceil \log[(1-\tau)^{|\mathbb{V}|}] * \theta \rceil$ if and only if $\mathbb{I}$ is an influential node set and every node is influenced by nodes in $\mathbb{I}$ no less than $\tau$.

[4]If there is a tie on the $f(\mathbb{I})$ value, we use the node ID to break the tie.

Next, we analyze the performance ratio of the proposed greedy algorithm MINS-GREEDY. As been well known, submodular functions and greedy algorithms have close relationship. Hence, we first introduce some theoretical results about submodular functions and the submodular cover problem. Consider a ground set $\mathbb{V}$ and a real function $f: 2^{\mathbb{V}} \to R$. $f(\cdot)$ is *submodular* if $\forall \mathbb{X} \subseteq \mathbb{Y} \subseteq \mathbb{V}$ and $x \in \mathbb{V} \backslash \mathbb{Y}$, $f(\mathbb{X} \cup \{x\}) - f(\mathbb{X}) \geq f(\mathbb{Y} \cup \{x\}) - f(\mathbb{Y})$. A function $f(\cdot)$ is called a *polymatroid function* if $f(\cdot)$ is submodular and increasing with $f(\emptyset) = 0$. Suppose $f(\cdot)$ is a polymatroid function on $2^{\mathbb{V}}$. A set $\mathbb{X} \subseteq \mathbb{V}$ is said to be a *submodular cover* of $(\mathbb{V}, f(\cdot))$ if $f(\mathbb{X}) = f(\mathbb{V})$. Moreover, consider that $f(\cdot)$ and a cost function $c(\cdot)$ are polymatroid functions on $2^{\mathbb{V}}$. The minimization problem $\min\{c(\mathbb{X}) \mid f(\mathbb{X}) = f(\mathbb{V}), \mathbb{X} \subseteq \mathbb{V})\}$ is called the *Minimum Submodular Cover with Submodular Cost* (MSC/SC) problem, where $c(\mathbb{X})$ is the cost of set $\mathbb{X}$. It is worth to mention that given any influential node set $\mathbb{I} \subseteq \mathbb{V}$, $f(\mathbb{I}) = -\theta * \log((1-\tau)^{|\mathbb{V}|}) = f(\mathbb{V})$. Let the cost function $c(\mathbb{X}) = |\mathbb{X}|$. Then the MINS selection problem can be formulated as: $\min\{c(\mathbb{I}) \mid f(\mathbb{I}) = f(\mathbb{V}), \mathbb{I} \subseteq \mathbb{V})\}$. Since $c(\mathbb{I})$ is linear (*i.e.*, modular), the MINS selection problem is a MSC/SC problem. Then, we have the following theorem [21], which is helpful when analyzing MINS-GREEDY.

THEOREM 3. [21] Suppose $f(\cdot)$ is a polymatroid function on $2^{\mathbb{V}}$, and $f(\mathbb{V}) \geq opt$ where $opt$ is the cost of a minimum submodular cover. For a greedy algorithm, if the selected $x$ in each iteration always satisfied that $\frac{\Delta_x f(\mathbb{X})}{c(x)} \geq 1$, then the greedy solution is a $1 + \rho \ln(\frac{f(\mathbb{V})}{opt})$-approximation, where $\rho$ is the curvature of the submodular cost $c$, *i.e.*, $\rho = \min_{\mathbb{Y} \subseteq \mathbb{V}} \frac{\sum_{y \in \mathbb{Y}} c(y)}{c(\mathbb{Y})}$. If $c$ is linear (*i.e.*, modular), then $\rho = 1$.

In the following, we will employ Theorem 3 to analyze the performance of MINS-GREEDY. First, we show the submodularity of $\mathbb{I}$ in Lemma 2.

LEMMA 2. $f(\mathbb{I})$ is a submodular function.

PROOF. Based on the definition of a submodular function, it is sufficient to prove that, for arbitrary two influential node sets $\mathbb{S}$ and $\mathbb{T}$, if $\mathbb{S} \subseteq \mathbb{T}$, we have for $\forall u_j \in \mathbb{V} \setminus \mathbb{T}$, $f(\mathbb{S} \cup \{u_j\}) - f(\mathbb{S}) \geq f(\mathbb{T} \cup \{u_j\}) - f(\mathbb{T})$.

Since $\mathbb{S} \subseteq \mathbb{T}$, $\forall u_i \in \mathbb{V}$, we have $\varrho_{u_i}^{\mathbb{S}} \geq \varrho_{u_i}^{\mathbb{T}}$ and $\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} \geq \varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}$. To prove $f(\mathbb{I})$ is a submodular function, we first prove that when $|\mathbb{T}| - |\mathbb{S}| = 1$, $f(\mathbb{S} \cup \{u_j\}) - f(\mathbb{S}) \geq f(\mathbb{T} \cup \{u_j\}) - f(\mathbb{T})$. Then, we extend it to the general case where $|\mathbb{T}| - |\mathbb{S}| = w > 1$, and the lemma still holds.

First, suppose $\mathbb{S} = \{u_1, u_2, \cdots, u_k\}$, then we have $\mathbb{S} \cup \{u_j\} = \{u_1, u_2, \cdots, u_k, u_j\}$. Let $\mathbb{T} = \{u_1, u_2, \cdots, u_k, u_{k+1}\}$, then we have $\mathbb{T} \cup \{u_j\} = \{u_1, u_2, \cdots, u_k, u_{k+1}, u_j\}$. Let

$$\Delta_{u_j} f(\mathbb{S}, u_i) = -\log(\max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau)) - (-\log(\max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau))) = \log(\frac{\max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau)}),$$

and

$$\Delta_{u_j} f(\mathbb{T}, u_i) = -\log(\max(\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}, 1-\tau)) - (-\log(\max(\varrho_{u_i}^{\mathbb{T}}, 1-\tau))) = \log(\frac{\max(\varrho_{u_i}^{\mathbb{T}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}, 1-\tau)}).$$

Furthermore, let

$$\begin{aligned}
\Delta_{u_j} f(\mathbb{S}) &= f(\mathbb{S} \cup \{u_j\}) - f(\mathbb{S}) \\
&= -\theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau)] + \theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau)] \\
&= \theta * \sum_{u_i \in \mathbb{V}} [-\log(\max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau))] + \theta * \sum_{u_i \in \mathbb{V}} [\log(\max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau))] \\
&= \theta * \sum_{u_i \in \mathbb{V}} [\log(\frac{\max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau)})] \\
&= \theta * \sum_{u_i \in \mathbb{V}} (\Delta_{u_j} f(\mathbb{S}, u_i)),
\end{aligned}$$

and

$$\Delta_{u_j} f(\mathbb{T}) = f(\mathbb{T} \cup \{u_j\}) - f(\mathbb{T}) = \theta * \sum_{u_i \in \mathbb{V}} (\Delta_{u_j} f(\mathbb{T}, u_i)).$$

To prove that $f(\mathbb{I})$ is a submodular function, we divide all possibilities into the following six cases:

1) For $\forall u_i \in \mathbb{V}$, which is influenced by $\mathbb{S}$, $\mathbb{S} \cup \{u_j\}$, $\mathbb{T}$, and $\mathbb{T} \cup \{u_j\}$, *i.e.*, $1 - \varrho_{u_i}^{\mathbb{S}} \geq \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} \geq \tau$, $1-\varrho_{u_i}^{\mathbb{T}} \geq \tau$, and $1-\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \geq \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\max(\varrho_{u_i}^{\mathbb{S}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}, 1-\tau)}) = \log(\frac{1-\tau}{1-\tau}) = 0$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{\max(\varrho_{u_i}^{\mathbb{T}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}, 1-\tau)}) = \log(\frac{1-\tau}{1-\tau}) = 0$. Hence, $\Delta_{u_j} f(\mathbb{S}, u_i) = \Delta_{u_j} f(\mathbb{T}, u_i)$.

2) For $\forall u_i \in \mathbb{V}$, which is influenced by $\mathbb{S} \cup \{u_j\}$, $\mathbb{T}$, and $\mathbb{T} \cup \{u_j\}$, however, not influenced by $\mathbb{S}$, *i.e.*, $1-\varrho_{u_i}^{\mathbb{S}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} \geq \tau$, $1 - \varrho_{u_i}^{\mathbb{T}} \geq \tau$, and $1 - \varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \geq \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{1-\tau}) > 0$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{1-\tau}{1-\tau}) = 0$. Hence, $\Delta_{u_j} f(\mathbb{S}, u_i) > \Delta_{u_j} f(\mathbb{T}, u_i)$.

3) For $\forall u_i \in \mathbb{V}$, which is influenced by $\mathbb{S} \cup \{u_j\}$, and $\mathbb{T} \cup \{u_j\}$, however, not influenced by $\mathbb{S}$, and $\mathbb{T}$, *i.e.*, $1 - \varrho_{u_i}^{\mathbb{S}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} \geq \tau$, $1 - \varrho_{u_i}^{\mathbb{T}} < \tau$, and $1 - \varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \geq \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{1-\tau})$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{T}}}{1-\tau})$. Since $\varrho_{u_i}^{\mathbb{S}} \geq \varrho_{u_i}^{\mathbb{T}} > 1 - \tau$, $\Delta_{u_j} f(\mathbb{S}, u_i) \geq \Delta_{u_j} f(\mathbb{T}, u_i)$.

4) For $\forall u_i \in \mathbb{V}$, which is influenced by $\mathbb{T}$, and $\mathbb{T} \cup \{u_j\}$, however, not influenced by $\mathbb{S}$, and $\mathbb{S} \cup \{u_j\}$, *i.e.*, $1 - \varrho_{u_i}^{\mathbb{S}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{T}} \geq \tau$, and $1 - \varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \geq \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}}) = \log(\frac{1}{1-p_{ij}}) > 0$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{1-\tau}{1-\tau}) = 0$. Hence, $\Delta_{u_j} f(\mathbb{S}, u_i) > \Delta_{u_j} f(\mathbb{T}, u_i)$.

5) For $\forall u_i \in \mathbb{V}$, which is influenced by $\mathbb{T} \cup \{u_j\}$, however, not influenced by $\mathbb{S}$, $\mathbb{S} \cup \{u_j\}$, and $\mathbb{T}$, *i.e.*, $1 - \varrho_{u_i}^{\mathbb{S}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{T}} < \tau$, and $1 - \varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \geq \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}})$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{T}}}{1-\tau})$. Then, we obtain $\Delta_{u_j} f(\mathbb{S}, u_i) - \Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}}) - \log(\frac{\varrho_{u_i}^{\mathbb{T}}}{1-\tau}) = \log(\frac{1}{1-p_{ij}}\frac{1-\tau}{\varrho_{u_i}^{\mathbb{T}}}) = \log(\frac{1-\tau}{\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}})$. Since $\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} \leq 1-\tau$, we obtain $\Delta_{u_j} f(\mathbb{S}, u_i) - \Delta_{u_j} f(\mathbb{T}, u_i) \geq 0$. Hence $\Delta_{u_j} f(\mathbb{S}, u_i) \geq \Delta_{u_j} f(\mathbb{T}, u_i)$.

6) For $\forall u_i \in \mathbb{V}$, which is not influenced by $\mathbb{S}$, $\mathbb{S} \cup \{u_j\}$, $\mathbb{T}$, or $\mathbb{T} \cup \{u_j\}$, *i.e.*, $1 - \varrho_{u_i}^{\mathbb{S}} < \tau$, $1 - \varrho_{u_i}^{\mathbb{S}\cup\{u_j\}} < \tau$, $1-\varrho_{u_i}^{\mathbb{T}} < \tau$, and $1-\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}} < \tau$, we have $\Delta_{u_j} f(\mathbb{S}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{S}}}{\varrho_{u_i}^{\mathbb{S}\cup\{u_j\}}}) = \log(\frac{1}{1-p_{ij}})$, and $\Delta_{u_j} f(\mathbb{T}, u_i) = \log(\frac{\varrho_{u_i}^{\mathbb{T}}}{\varrho_{u_i}^{\mathbb{T}\cup\{u_j\}}}) = \log(\frac{1}{1-p_{ij}})$. Hence, $\Delta_{u_j} f(\mathbb{S}, u_i) = \Delta_{u_j} f(\mathbb{T}, u_i)$.

In summary, $\Delta_{u_j} f(\mathbb{S}, u_i) \geq \Delta_{u_j} f(\mathbb{T}, u_i)$ in all the cases. Therefore,

$$\Delta_{u_j} f(\mathbb{S}) = \sum_{u_i \in \mathbb{V}} (\Delta_{u_j} f(\mathbb{S}, u_i)) \geq \sum_{u_i \in \mathbb{V}} (\Delta_{u_j} f(\mathbb{T}, u_i)) = \Delta_{u_j} f(\mathbb{T}).$$

Now, suppose $\mathbb{S} = \{u_1, u_2, \cdots, u_k\}$, and $\mathbb{T} = \{u_1, u_2, \cdots, u_k, u_{k+1}, \cdots, u_{k+w}\}$. We obtain

$$\begin{aligned}
f(\mathbb{S} \cup \{u_j\}) - f(\mathbb{S}) &\geq f(\mathbb{S} \cup \{u_{k+1}, u_j\}) - f(\mathbb{S} \cup \{u_{k+1}\}) \\
&\geq f(\mathbb{S} \cup \{u_{k+1}, u_{k+2}, u_j\}) - f(\mathbb{S} \cup \{u_{k+1}, u_{k+2}\}) \\
&\geq f(\mathbb{S} \cup \{u_{k+1}, u_{k+2}, \cdots, u_w, u_j\}) - f(\mathbb{S} \cup \{u_{k+1}, u_{k+2}, \cdots, u_w\}) \\
&= f(\mathbb{T} \cup \{u_j\}) - f(\mathbb{T}).
\end{aligned}$$

Therefore, $f(\mathbb{I})$ is a submodular function. □

Now, we can make the following conclusion.

THEOREM 4. $f(\mathbb{I})$ is a polymatroid function on $2^{\mathbb{V}}$.

PROOF. According to Lemma 1 and Lemma 2, $f(\mathbb{I})$ is an increasing, submodular function with $f(\emptyset) = 0$. Hence, we conclude that $f(\mathbb{I})$ is a polymatroid function on $2^{\mathbb{V}}$. □

Before employing Theorem 3 to analyze the performance ratio of MINS-GREEDY, we give the following important lemmas first.

LEMMA 3. In MINS-GREEDY shown in Algorithm 1, if $f(\mathbb{I}) < -\theta * \log[(1-\tau)^{|\mathbb{V}|}]$, then there exists a node $u_k \in \mathbb{V} \setminus \mathbb{I}$ such that $f(\mathbb{I} \cup \{u_k\}) > f(\mathbb{I})$.

PROOF. Let $\mathbb{I}_t$ be the selected influential node set after the $t$-th iteration. In Algorithm 1, if $f(\mathbb{I}_t) < -\theta * \log[(1-\tau)^{|\mathbb{V}|}]$, the algorithm continues. Therefore, there must exists a node $u_j \in \mathbb{V} \setminus \mathbb{I}$ satisfying $\varrho_{u_j}^{\mathbb{I}_t} > 1-\tau$. Furthermore, let $u_k$ be the node selected to add to $\mathbb{I}_t$ in step 3 of Algorithm 1 during the $(t+1)$-th iteration. Since $u_k$ maximizes the $f(\cdot)$ value based on Algorithm 1, we have

$$\begin{aligned}
f(\mathbb{I}_{t+1}) - f(\mathbb{I}_t) &= f(\mathbb{I}_t \cup \{u_k\}) - f(\mathbb{I}_t) \\
&\geq f(\mathbb{I}_t \cup \{u_j\}) - f(\mathbb{I}_t) \\
&= -\theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)] + \\
&\quad \theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{I}_t}, 1-\tau)] \\
&= \theta * \sum_{u_i \in \mathbb{V}} -\log(\max(\varrho_{u_i}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)) + \\
&\quad \theta * \sum_{u_i \in \mathbb{V}} \log(\max(\varrho_{u_i}^{\mathbb{I}_t}, 1-\tau)) \\
&= \theta * \sum_{u_i \in \mathbb{V}} \log(\frac{\max(\varrho_{u_i}^{\mathbb{I}_t}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)}) \\
&= \theta * [\sum_{u_i \in \mathbb{V} \setminus \{u_j\}} \log(\frac{\max(\varrho_{u_i}^{\mathbb{I}_t}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)}) + \\
&\quad \log(\frac{\max(\varrho_{u_j}^{\mathbb{I}_t}, 1-\tau)}{\max(\varrho_{u_j}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)})] \\
&= \theta * [\sum_{u_i \in \mathbb{V} \setminus \{u_j\}} \log(\frac{\max(\varrho_{u_i}^{\mathbb{I}_t}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{I}_t \cup \{u_j\}}, 1-\tau)}) + \\
&\quad \log(\frac{\varrho_{u_i}^{\mathbb{I}_t}}{1-\tau})] > 0.
\end{aligned}$$

Therefore, we have $f(\mathbb{I}_t \cup \{u_k\}) > f(\mathbb{I}_t)$. □

LEMMA 4. *1)* $f(\mathbb{V}) \geq opt$, where $opt = c(\mathbb{I}_{opt}) = |\mathbb{I}_{opt}|$ is the cost of the optimal solution of the MINS selection problem denoted by $\mathbb{I}_{opt}$. *2)* The selected $u_i$ of each iteration of MINS-GREEDY shown in Algorithm 1 satisfies that $\frac{\Delta_{u_i} f(\mathbb{I})}{c(u_i)} \geq 1$, where $\Delta_{u_i} f(\mathbb{I}) = f(\mathbb{I} \cup \{u_i\}) - f(\mathbb{I})$.

PROOF. For 1), Based on Definition 4.1, we know $\theta \geq 1/(-\log(1-\tau))$. Therefore,

$$f(\mathbb{V}) = -\theta * \log[(1-\tau)^{|\mathbb{V}|}] \geq |\mathbb{V}| \geq |\mathbb{I}_{opt}| = opt.$$

For 2), According to Lemma 3, we know that if $f(\mathbb{I}) < -\theta * \log[(1-\tau)^{|\mathbb{V}|}]$, then there exists a node $u_k \in \mathbb{V} \setminus \mathbb{I}$ such that $f(\mathbb{I} \cup \{u_k\}) > f(\mathbb{I})$. In other words, if Algorithm 1 does not terminate, there must exist a node $u_j \in \mathbb{V}$ satisfying $\varrho_{u_j}^{\mathbb{I}} > 1-\tau$, and

$$\begin{aligned}
\Delta_{u_k} f(\mathbb{I}, u_j) &= -\log(\max(\varrho_{u_j}^{\mathbb{I} \cup \{u_k\}}, 1-\tau)) - \\
&\quad (-\log(\max(\varrho_{u_j}^{\mathbb{I}}, 1-\tau)) \\
&= \log(\frac{\max(\varrho_{u_j}^{\mathbb{I}}, 1-\tau)}{\max(\varrho_{u_j}^{\mathbb{I} \cup \{u_k\}}, 1-\tau)}) > 0.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\Delta_{u_k} f(\mathbb{I}) &= f(\mathbb{I} \cup \{u_k\}) - f(\mathbb{I}) \\
&= -\theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{I} \cup \{u_k\}}, 1-\tau)] + \\
&\quad \theta * \log[\prod_{u_i \in \mathbb{V}} \max(\varrho_{u_i}^{\mathbb{I}}, 1-\tau)] \\
&= \theta * \sum_{u_i \in \mathbb{V}} [-\log(\max(\varrho_{u_i}^{\mathbb{I} \cup \{u_k\}}, 1-\tau))] + \\
&\quad \theta * \sum_{u_i \in \mathbb{V}} [\log(\max(\varrho_{u_i}^{\mathbb{I}}, 1-\tau))] \\
&= \theta * \sum_{u_i \in \mathbb{V}} [\log(\frac{\max(\varrho_{u_i}^{\mathbb{I}}, 1-\tau)}{\max(\varrho_{u_i}^{\mathbb{I} \cup \{u_k\}}, 1-\tau)})] \\
&= \theta * \sum_{u_i \in \mathbb{V}} (\Delta_{u_k} f(\mathbb{I}, u_i)).
\end{aligned}$$

To prove the lemma, we divide the possibilities into the following two cases:

1) $u_j$ is not influenced by $\mathbb{I} \cup \{u_k\}$. Then,

$$\Delta_{u_k} f(\mathbb{I}, u_j) = \log(\frac{\varrho_{u_j}^{\mathbb{I}}}{\varrho_{u_j}^{\mathbb{I} \cup \{u_k\}}}) = \log(\frac{1}{1-p_{kj}}).$$

Since we know that $\theta \geq \frac{1}{(\log \frac{1}{\max_{p_{ij}<1}(1-p_{ij})})}$. Hence

$$\Delta_{u_k} f(\mathbb{I}) = \theta * \sum_{u_j \in \mathbb{V}} (\Delta_{u_k} f(\mathbb{I}, u_j)) \geq \theta * \log(\frac{1}{1-p_{kj}}) \geq 1.$$

2) $u_j$ is influenced by $\mathbb{I} \cup \{u_k\}$. Then,

$$\Delta_{u_k} f(\mathbb{I}, u_j) = \log(\frac{\varrho_{u_j}^{\mathbb{I}}}{1-\tau}) = \log(\frac{\varrho_{u_j}^{\mathbb{I}}}{1-\tau}).$$

Since $\theta \geq \frac{1}{(\log \frac{\min\{\prod_{u_j \in \mathbb{S}}(1-p_{ij}) \mid u_j \in \mathbb{V}, \mathbb{S} \subseteq \mathbb{V}, \prod_{u_j \in \mathbb{S}}(1-p_{ij}) > 1-\tau\}}{1-\tau})}$, we have

$$\Delta_{u_k} f(\mathbb{I}) = \theta * \sum_{u_j \in \mathbb{V}} (\Delta_{u_k} f(\mathbb{I}, u_j)) \geq \theta * \log(\frac{\varrho_{u_j}^{\mathbb{I}}}{1-\tau}) \geq 1.$$

In summary, $\frac{\Delta_{u_i} f(\mathbb{I})}{c(u_i)} \geq \frac{\Delta_{u_k} f(\mathbb{I})}{c(u_i)} = \Delta_{u_k} f(\mathbb{I}) \geq 1$. □

Now, we are ready to analyze the performance ratio of Algorithm 1 as follows.

THEOREM 5. The performance ratio of the greedy algorithm shown in Algorithm 1 is $1 + \ln(\frac{-\theta|\mathbb{V}|\log(1-\tau)}{opt})$, where $opt$ is the size of the optimal solution of MINS.

PROOF. According to Theorem 4, our proposed contribution function $f(\mathbb{I})$ is a polymatroid function on $2^{\mathbb{V}}$. Moreover, based on Lemma 4, and Theorem 3, MINS-GREEDY produces an approximation solution with a factor of $1 + \ln(\frac{f(\mathbb{V})}{opt}) = 1 + \ln(\frac{-\theta|\mathbb{V}|\log(1-\tau)}{opt})$ from the optimal, where $opt$ is the size of the optimal solution of MINS. □

# 5. PERFORMANCE EVALUATION

Since there are no existing works studying the MINS selection problem under the independent cascade model currently, in the simulations and experiments, the results of MINS-GREEDY (denoted by MINS) are compared with the most related work [7] denoted by PIDS, and the optimal solution

of MINS, which are obtained by exhausting searching, denoted by OPTIMAL. To ensure fairness of comparisons, the condition of termination to the algorithm proposed in [7] is changed to find a PIDS, such that every node in the network is influenced no less than the same threshold $\tau$ in MINS.

## 5.1 Simulation Results

### 5.1.1 Simulation Setting

We build our own simulator to generate random graphs based on the random graph model $G(n,p) = \{G \mid G$ has $n$ nodes and an edge between any pair of nodes is generated with probability $p\}$. For $G = (V,E) \in G(n,p)$, $u_i$, $u_j \in V$, and $(u_i,u_j) \in E$, the associated social influence $0 < p_{ij} \leq 1$ is randomly generated. For each specific setting, 100 instances are generated. The results are the average values of these 100 instances. In the following, we show the simulation results under different scenarios.

### 5.1.2 Simulation Results in Random Graphs

The objectives of MINS and PIDS are both to minimize the size of the constructed subsets. In this subsection, we check the size of the solutions of MINS, PIDS and OPTIMAL under different scenarios in random graphs. In this simulation, we consider the following tunable parameters: the network size $n$, the probability $p$ to create an edge in the random graph model $G(n,p)$, and the user pre-defined influence threshold $\tau$. Since we adopt exhaust searching to find the optimal solution of MINS, it is impractical to test on large scale networks. Hence, we first run a set of simulations on small scale networks of network size changing from 10 to 20, and the results are shown in Fig. 3.

The impacts of $n$, $p$, and $\tau$ on the size of the solutions of MINS, PIDS, and OPTIMAL are shown in Fig. 3(a), (b), and (c), respectively. From Fig. 3(a), we can see that the sizes of the solutions of all the three algorithms increase when $n$ increases. This is because more nodes need to be influenced when network size increases. Additionally, for a specific network size, PIDS produces a larger sized solution than MINS. This is because MINS tries to find the most influential node of the network (which has the largest $f(\mathbb{I})$ value) in each iteration, while PIDS gives the node with the largest degree the highest priority instead. However, a large degree does not necessarily implies high influence to a social network. Furthermore, we can see that the size of the MINS solution is very close to the optimal result. On average, MINS produces 0.62 more nodes than the optimal solution, while PIDS produces 3.53 more nodes than the optimal solution, which implies that our proposed greedy algorithm MINS-GREEDY can produce a close approximation solution to the optimal solution.

From Fig. 3(b), we can see that the solution sizes of all the three algorithms decrease when $p$ increases. This is because that a large $p$ means more edges in the network, so that one selected influential node may influence more nodes. Again, for a specific $p$, PIDS produces a larger sized solution than MINS. This is because the objective of PIDS is not aimed to obtain the most influential nodes in the network. MINS again can construct a solution with similar size of the optimal solution. On average, MINS only produces 8.07% more nodes than the optimal solution, while PIDS produces 24.30% more nodes than the optimal solution.

From Fig. 3(c), we can see that the sizes of the solutions of all the algorithms increase when $\tau$ increases, since large $\tau$ value means that more nodes need to be put into the initial active node set to influence all the other nodes. Furthermore, MINS has similar performance with optimal, and has a better performance than PIDS since the greedy criterion of PIDS is the node with the highest degree first. On average, MINS produces 5.01% more nodes than the optimal solution, while PIDS produces 24.44% more nodes than the optimal solution. This reason is similar as we mentioned before.

Additionally, we run a set of simulations on large scale networks of network size changing from 100 to 1000. The impacts of $n$, $p$, and $\tau$ on MINS and PIDS are shown in Fig. 4. From Fig. 4(a), we can see that the solution sizes of MINS and PIDS are both increase when $n$ increases. This is because that more influential nodes are required for large social networks. Moreover, MINS can find an influential node set that is much smaller than that of PIDS, since MINS tries to find the most influential node in the network during each iteration. On average, MINS produces an influential node set of size 46.67% less than PIDS.

From Fig. 4(b), we can see that the solution sizes of PIDS and MINS are both decrease when $p$ increases. $p$ increases means the number of edges in the network increases, which further implies that the average number of neighbors of each node increases. Hence, one selected influential node may influence more nodes when $p$ increases. For a specific $p$, PIDS again produces larger sized solution than MINS. On average, PIDS produces 21.6% more nodes than MINS. Additionally, the decreasing trend of PIDS is fast when $p$ increases. This is because when $p$ is small, the expected degree of all nodes is small. Hence, PIDS may find a solution through many iterations till it find a solution satisfying that every node in the network is influenced by the solution no less than $\tau$. When $p$ is large, larger degree nodes could be added into the solution first, so that PIDS might terminate quicker and followed by an influential node set of small size.

From Fig. 4(c), because of similar reasons as in analyzing Fig. 3(c), we can see that the solution sizes of PIDS and MINS increase when $\tau$ increases. Moreover, PIDS outputs more nodes than MINS. On average, PIDS produces 6.15 more nodes than that of MINS.

## 5.2 Experimental Results

### 5.2.1 Experimental Setting

We also perform our experiments on a real-world data set: *academic coauthor network*, which is extracted from the academic search system Arnetminer [22]. The coauthor data set consists of $640,134$ authors and $1,554,643$ coauthor relations [2]. The testing data sets used in this experiment are extracted from the coauthor data set. In this experiment, the network size $n$, which is the number of authors, is change from 100 to 1100. The social influence of each pair of nodes is calculated using the methodology introduced in [2]. Fig. 5 summarizes the distribution of social influence values among the authors in one testing data set. From Fig. 5, we can see that most of the social influence values fall into the range $0 < p_{ij} < 0.05$, which are very small. Moreover, the largest social influence value in the test data set is close to 0.20. Based on this observation, we let $\tau$ change from 0.005 to 0.05 in this experiment. For each specific setting, 100

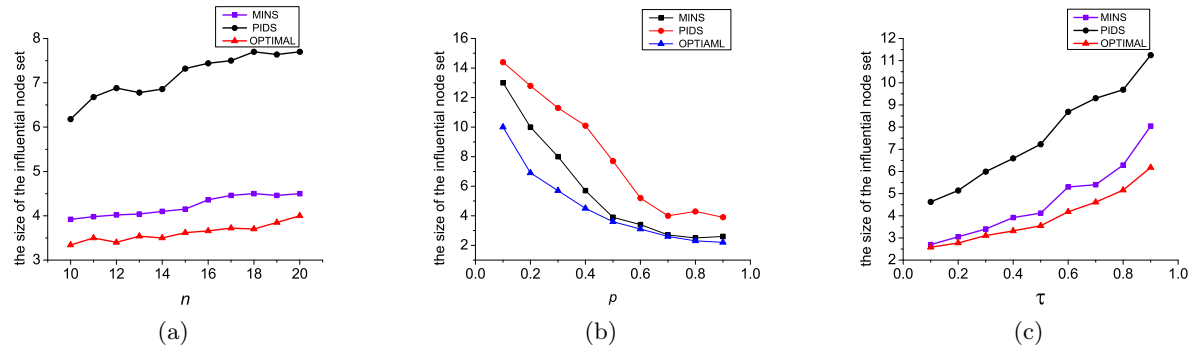



**Figure 3: The size of solutions on small scale networks. The default setting are $n = 15$, $p = 0.5$, and $\tau = 0.5$.**

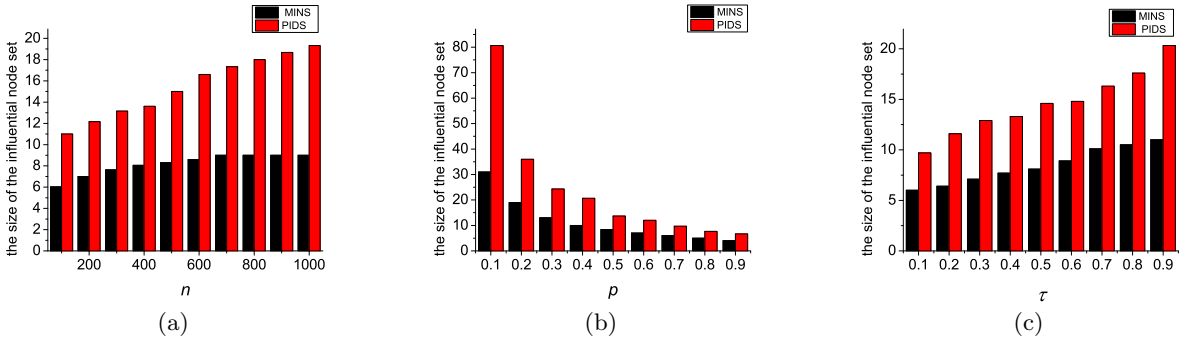



**Figure 4: The size of solutions on large scale networks: The default settings are $n = 500$, $p = 0.5$, and $\tau = 0.5$.**

instances are generated. The results are the average values of these 100 instances.

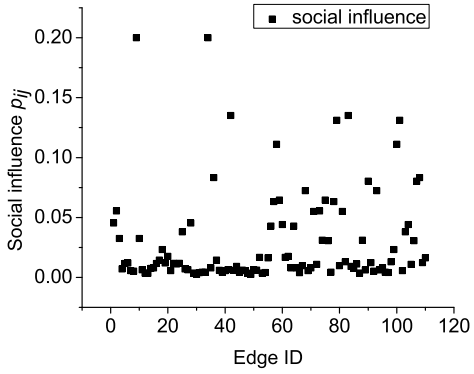



**Figure 5: The distribution of social influences of the coauthor data set.**

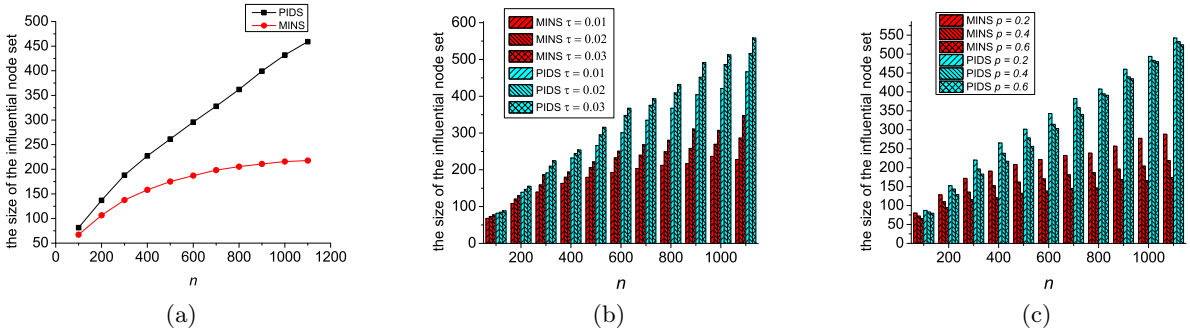
### 5.2.2 *Experiments on Arnetminer Data*

For $\tau = 0.008$, the impacts of $n$ on the size of the solutions of MINS and PIDS are shown in Fig. 6(a). From Fig. 6(a), the solution sizes of PIDS and MINS increase when $n$ increases. This is because, when the network becomes large, more influential nodes are requireed to influence the whole network. Furthermore, because of similar reasons analyzed before, MINS produces smaller influence node sets than PIDS. This is consistent with the simulation results. On average, MINS selects 23.49% less influential nodes than that of PIDS.

The impacts of $\tau$ on MINS and PIDS are shown in Fig. 6(b). From Fig. 6(b), we can see that the solution sizes of both algorithms increase when $\tau$ increases. The reason is obvious, since in order to guarantee the requirement that $\forall u_i \in \mathbb{V}, 1 - \prod_{u_j \in \mathbb{A}^{\mathbb{I}}(u_i)} (1 - p_{ij}) \geq \tau$, more nodes need to be selected as the initial active nodes in both MINS and PIDS. Again, PIDS selects more influence nodes than MINS, since it takes node degree as its greedy criteria instead of node influence. On average, MINS selects 18.45% less influential nodes than PIDS.

We also perform an experiment on a synthesized data set. For the data set used in Section 5.2.1, we randomly generate additional non-redundant $p$ percent of existing edges to the original testing data set. The impacts of $p$ on both algorithms are shown in Fig. 6(c). From Fig. 6(c), we can see that the solution sizes of PIDS and MINS decrease when $p$ increases, since large $p$ value means more edges in the network. In other words, one node may influence more nodes when $p$ increase. Hence, smaller sized initial active node sets may satisfy the requirements both for PIDS and MINS. Again, the size of PIDS's solution is larger than that of MINS's solution, which is also consistent with the simulation results. On average, MINS selects 24.80% less influential nodes than PIDS.

From the results of simulations on the random graphs, and the results of experiments on the real world data sets, we can conclude that the size of the constructed initial active node set of MINS is smaller than it of PIDS. Moreover, the




(a) (b) (c)

**Figure 6: The size of solutions on real-world data sets.**

solution of MINS is very close to the optimal solutions of MINS in small scale networks.

# 6. CONCLUSION

In this paper, we study the Minimum-sized Influential Node Set (MINS) selection problem which has useful commercial applications in social networks. We show by reduction that MINS is NP-hard under the independent cascade model. Subsequently, a greedy algorithm called MINS-GREEDY is proposed to solve the problem, followed by the theoretical analysis of its performance ratio. Furthermore, we validate our proposed algorithm through simulations on random graphs, and experiments on real world data sets. The simulation results indicate that MINS-GREEDY can construct smaller sized satisfied initial active node sets than the most related work PIDS. Moreover, MINS-GREEDY has very close performance to the optimal solution of MINS in small scale networks.

## Acknowledgment

This research is partly supported by the National Science Foundation (NSF) under Grants Nos. CNS-1152001, CNS-1252292, and by the Kennesaw State University College of Science and Mathematics the Interdisciplinary Research Opportunities (IDROP) Program.

# 7. REFERENCES

[1] K. Saito, R. Nakana, and M. Kimura, *Prediction of Information Diffusion Probabilities for Independent Cadcade Model*, KES'08.

[2] J. Tang, J. Sun, C. Wang, and Z. Yang, *Social Influence Analysis in Large-Scale Networks*, KDD'09.

[3] A. Goyal, F. Bonchi, L. Laskhmanan, *Learning Influence Probabilities in Social Networks*, WSDM'10.

[4] C. Wang, J. Tang, J. Sun, and J. Han, *Dynamic Social Influence Analysis through Time-Dependent Factor Graphs*, ASONAM'11.

[5] F. Wang, E. Camacho, and K. Xu, *Positive Influence Dominating Set in Online Social Networks*, Lecture Notes in Computer Science, 2009.

[6] D. Kempe, J. Kleinberg, and E. Tardos, *Maximizing the Spread of Influence through a Social Network*, KDD'03.

[7] F. Wang, H. Du, E. Camacho, K. Xu, W. Lee, Yan, Shi, and S. Shan, *On Positive Influence Dominating Sets in Social Networks*, TCS, 2011.

[8] P. Domingos, and M. Richardson, *Mining the network value of customers*, KDD'01.

[9] M. Richardson, and P. Domingos, *Mining Knowledge-sharing sites for viral marketing*, KDD'02.

[10] D. Kempe, J. Kleinberg, and E. Tardos, *Influential Nodes in a Diffusion Model for Social Networks*, ICALP'05.

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriesen, and N. Glance, *Cost-effective outbreak detection in networks*, KDD'07.

[12] W. Chen, Y. Yuan, and S. Yang, *Efficient Influence Maximization in Social Networks*, SIGKDD, pp. 199-208, 2010.

[13] W. Chen, Y. Yuan, and L. Zhang, *Scalable Influence Maximization in Social Networks under the Linear Threshold Model*, KDD'10.

[14] W. Chen, C. Wang, and Y. Wang, *Scalable influence maximization for prevalent viral marketing in large-scale social networks*, KDD'10.

[15] A. Goyal, F. Bonchi, and L. Lakshmanan, *A data-based approach to social influence maximization*, VLDB, 5(1):73-84, 2011.

[16] T.N. Dinh, Y. Shen, D.T. Nguyen, and M.T. Thai, *Cost-effective Viral Marketing for Time-critical Campaigns in Large-scale Social Networks*, ToN, Volume:PP, Issue:99, 2013.

[17] F. Zou, Z. Zhang, and W. Wu, *Latency-Bounded Minimum Influential Node Selection in Social Networks*, WASA'09.

[18] F. Zou, J. Willson, Z. Zhang, W. Wu, *Fast Information Propagation in Social Networks*, DMAA, 2010.

[19] J. He, S. Ji, X. Liao, H. M. Haddad, and R. Beyah, *Minimum-sized Positive Influential Node Set Selection for Social Networks: Considering Both Positive and Negative Influences*, IPCCC'13.

[20] X. Zhu, J. Yu, W. Lee, D. Kim, S. Shan, and D. Du, *New Dominating Sets in Social Networks*, JGO, 2010.

[21] P. Wan, D. Du, P. Pardalos, and W. Wu, *Greedy Approcimations for Minimum Submodular Cover with Submodular Cost*, Computer Optimization and Applications, 45:463-474, 2010.

[22] http://arnetminer.org